Darja Fišer
Univerza v Ljubljani, Filozofska fakulteta
darja.fiser@guest.arnes.si

UDK 81'374:811.163.6'373:004.6

PRISTOPI K IZDELAVI LEKSIKALNIH PODATKOVNIH ZBIRK

Prispevek predstavlja leksikalni podatkovni zbirki *WordNet* in *EuroWordNet*, ki temeljita na povezavi pojmov z leksikalnimi in semantičnimi relacijami. V zadnjem desetletju so na njuni podlagi izdelali podobne leksikalne zbirke za mnoge druge jezike, pri čemer so raziskovalci uporabili predvsem dva pristopa; gradnjo neodvisnih besednih mrež, ki so jih kasneje združili v večjezične leksikalne zbirke, in prevajanje vsebine leksikalnih zbirk v ciljni jezik s prevzemanjem relacij med posameznimi vnosi. Slednji pristop je priljubljen predvsem zaradi možnosti avtomatizacije procesa in poenotene strukture dobljenih zbirk, zato ga v prispevku preizkusimo za gradnjo slovenske besedne mreže in rezultate primerjamo z jezikovnomotivirano besedno mrežo, ki smo jo izdelali s pomočjo korpusa *Fida* in drugih eno- in večjezičnih virov za slovenščino.

1 Uvod

Mentalni leksikon, dinamična organizacija besed v naših mislih, je temelj človeških jezikovnih sposobnosti in je sestavljen iz obsežne in kompleksne mreže mentalnih reprezentacij, asociacij in procesov. Pri proučevanju mentalnega leksikona nas predvsem zanima, kako je leksika urejena in kako dostopamo do nje, sodobne modele mentalnega leksikona pa s pridom uporabljamo tudi v različnih aplikacijah za računalniško obdelavo naravnega jezika. Računalniški modeli se od mentalnega leksikona razlikujejo tako po organizacijski kot po vsebinski plati. Najpomembnejša razlika je, da mentalni leksikon ni in ne more biti zapisan. V primerjavi z mentalnim leksikonom so leksikalne zbirke veliko bolj omejene, njihova organizacija je močno poenostavljena, vsebina pa nespremenljiva in kmalu postane zastarela, prav tako je v zbirke vključenih tudi razmeroma malo informacij o določenem leksemu (Aitchison 2004: 10–15).

V prispevku predstavimo *WordNet* in *EuroWordNet*, najbolj razširjeni leksikalni zbirki v zadnjem desetletju, ki temeljita na povezavi pojmov z leksikalnimi in semantičnimi relacijami. Ta obsežna projekta sta v zadnjih letih pospešila razvoj podobnih leksikalnih zbirk tudi za številne druge jezike, na primer za bolgarski, češki, grški, romunski, turški in srbski jezik v okviru projekta *BalkaNet*,[1] pri čemer raziskovalci

[1] <http://www.ceid.upatras.gr/Balkanet> [15. 6. 2005].

največkrat uporabljajo dva pristopa: gradnjo neodvisnih besednih mrež z naknadnim združevanjem v večjezične leksikalne zbirke in razširitveni model, po katerem sinonimske nize besed (sinsete), ki predstavljajo posamezne leksikalne pojme (npr. *avto, avtomobil*), prevedejo v ciljni jezik in prevzamejo relacije med njimi.

Zaradi možnosti avtomatizacije procesa in enotne strukture izdelanih zbirk se vse bolj uveljavlja razširitveni model. Zato v nadaljevanju prispevka ta model preizkusimo za gradnjo slovenske besedne mreže za pomensko polje |sorodstvo|. Razpravo sklenemo z vrednotenjem rezultatov in s primerjavo razširitvene besedne mreže z jezikovnomotivirano, ki smo jo izdelali z referenčnim korpusom *Fida* in s pomočjo drugih različnih eno- in večjezičnih jezikovnih virov za slovenščino.

2 Sodobne leksikalne podatkovne zbirke

Leksikalne zbirke so odličen vir leksikalnega znanja za številne aplikacije računalniške obdelave naravnega jezika. Za različne potrebe so raziskovalci razvili številne leksikalne zbirke, med katerimi so najodmevnejše *Cyc*,[2] *EuroWordNet*,[3] *FrameNet*,[4] *HowNet*[5] in *WordNet*.[6] Medtem ko sta struktura in podrobna organizacija pojmov v posameznih zbirkah različna, si vse prizadevajo predstaviti hierarhijo jezikovno neodvisnih pojmov in uporabljajo podobne relacije za ustvarjanje povezav med posameznimi pojmi. V nadaljevanju predstavljamo osnovne lastnosti *WordNeta* in *EuroWordneta*.

2.1 *WordNet*

Začetki *WordNeta* segajo v osemdeseta leta minulega stoletja, ko je George A. Miller z Univerze v Princetonu začel preizkušati prvo relacijsko podatkovno zbirko za angleški jezik. Sčasoma se je razvila v najobsežnejšo podatkovno zbirko te vrste in danes jo strokovnjaki uporabljajo za najrazličnejše računalniške aplikacije, povezane z obdelavo naravnega jezika.

WordNet je leksikalna podatkovna zbirka, v kateri so samostalniki, glagoli, pridevniki in prislovi urejeni v sinonimske nize (**sinsete**), ki predstavljajo posamezne leksikalne pojme, npr. {avto, avtomobil}.[7] Sinsetom je dodana razlaga, pogosto tudi primer rabe in oznaka za področje, iz katerega izhaja, npr. *računalništvo*, med seboj pa so sinseti povezani z različnimi semantičnimi in leksikalnimi relacijami. **Semantične relacije**, kot so hipo- in hipernimija ter meronimija, povezujejo sinsete ({avto, avto-

[2] <http://www.opencyc.org> [15. 6. 2005].

[3] <http://www.illc.uva.nl/EuroWordNet> [15. 6. 2005].

[4] <http://framenet.icsi.berkeley.edu> [15. 6. 2005].

[5] <http://www.keenage.com/zhiwang/e_zhiwang.html> [15. 6. 2005].

[6] <http://wordnet.princeton.edu> [15. 6. 2005].

[7] Zaviti oklepaji ({}) označujejo sinsete, puščice (→) pa relacije med njimi.

mobil} → {vozilo}), **leksikalne relacije**, kot je protipomenskost, pa veljajo zgolj med posameznimi leksemi (*lep – grd*) (Fellbaum 1998: 3–17).

2.2 *EuroWordNet*

WordNet je v devetdesetih letih dvajsetega stoletja spodbudil razvoj večjezične leksikalne zbirke z imenom *EuroWordNet*, ki temelji na povezavi pojmov med številnimi evropskimi jeziki (češki, estonski, francoski, italijanski, nemški, nizozemski in španski).

EuroWordNet je prevzel *WordNetovo* strukturo, vendar je zaradi svoje večjezične razsežnosti sestavljen iz treh modulov. **Jezikovno odvisni modul** vsebuje pojme, zastopane v določenem jeziku, in je pravzaprav neodvisna besedna mreža za določen jezik. Vsi jezikovno odvisni moduli so med seboj povezani z **jezikovno neodvisnim**, imenovanim *interlingual index* (ILI). Preko ILI-ja je kateri koli pojem v nekem jeziku mogoče prevesti v drug jezik, ki je prav tako vključen v zbirko. Z **modulom osnovnih pojmov in ontologijo domen** pa je omogočena večja združljivost različnih jezikovno neodvisnih komponent, saj prispeva enotno hierarhično strukturo za najpomembnejše pojme in seznam področnih oznak (Vossen 2005: 5–11).

2.3 Uporabnost leksikalnih podatkovnih zbirk

Leksikalne zbirke z eksplicitno izraženimi semantičnimi relacijami med leksemi so zelo uporabne pri računalniški obdelavi naravnega jezika. S pomočjo semantičnih odnosov je mogoče meriti semantično razdaljo med leksemi, kar nam pomaga pri **avtomatskem razdvoumljanju** večpomenskih besed (ang. *automatic sense disambiguation*). Leksika je urejena v pomenska polja, ki računalniku nudijo širši kontekst za iskanje informacij o tem, kateri pomen besede je mišljen v določenem kontekstu.

Številni sistemi za **pridobivanje informacij** iz besedil (ang. *information retrieval*) temeljijo predvsem na statističnem ujemanju med besedami v poizvedbi in tistimi v besedilih v podatkovni zbirki. Priklic relevantnih besedil pa je s predstavljenimi leksikalnimi zbirkami mogoče izboljšati z identifikacijo dodatnih besed, ki bodo najverjetneje v istem sobesedilu kot iskane besede.

Glede na to, da so v *WordNetu* in sorodnih podatkovnih zbirkah besede urejene po posameznih pomenih, jih lahko z dodajanjem identifikacijske številke relevantnega sinseta s pridom izkoristimo tudi za **označevanje pomenov** besed (ang. *semantic tagging*) v besedilih in korpusih.

Če leksikalne podatkovne zbirke razumemo kot **ontologije**, ki služijo eksplicitni reprezentaciji za urejanje entitet in odnosov med njimi v določeni domeni, jih je mogoče vključiti v najrazličnejše sisteme za računalniško obdelavo naravnega jezika, ki temeljijo na znanju (Noy 2003).

Nenazadnje pa je *WordNet* kot podatkovna zbirka s splošnim besediščem tudi trdna osnova za kasnejši razvoj **terminoloških zbirk** z najrazličnejših strokovnih področij,

ki so nato kot samostojna orodja uporabna za številne aplikacije na ozko določenih področjih.

3 Gradnja novih leksikalnih podatkovnih zbirk

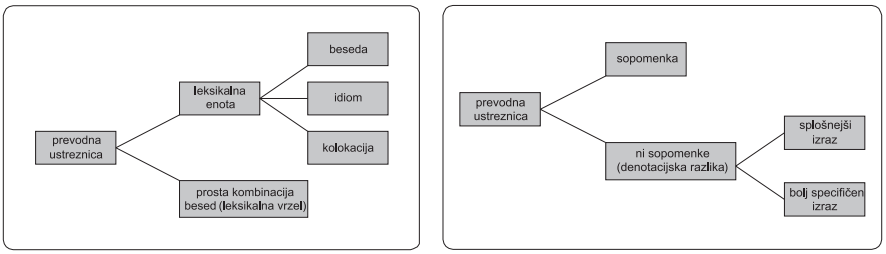
Za gradnjo novih, *WordNetu* podobnih leksikalnih zbirk sta najbolj razširjena dva pristopa. Po prvem so bile zgrajene zbirke za večino jezikov, vključenih v projekt *EuroWordNet*. Raziskovalne skupine so ob upoštevanju skupnih smernic neodvisno razvile zbirke za svoje jezike, nato pa so te neodvisne zbirke skušale združiti. Med **gradnjo neodvisnih podatkovnih zbirk** so strokovnjaki sprejeli številne subjektivne odločitve, ki so botrovale temu, da so se končane zbirke med seboj precej razlikovale, pri čemer razlogi za razlike največkrat niso bili jezikovni, temveč je do njih prihajalo zaradi različnih subjektivnih odločitev glede sestave zbirk. Naj pripomnimo, da se temu najbrž ni mogoče v celoti izogniti, kljub temu pa jih lahko precej omilimo s korpusno objektivizacijo jezikovnih podatkov.

Diskrepance med posameznimi zbirkami skuša čim bolj zmanjšati drugi pristop, ki zahteva strogo upoštevanje meril za gradnjo angleškega *WordNeta* in skupnih objektivnih odločitev. Leksikalne zbirke, zgrajene s tem pristopom, se skušajo čim bolj držati sinsetov in semantičnih relacij angleškega *WordNeta* po načelu: če med sinsetoma v angleškem *WordNetu* velja neka relacija, velja ista relacija tudi med ekvivalentnima sinsetoma v novem jeziku (Vossen 1998).

Vossen (1998) ta pristop imenuje »**razširitveni model**« (ang. *expand model*), za katerega trdi, da je veliko preprostejši od prvega in zagotavlja najvišjo možno stopnjo ujemanja med različnimi jeziki. Pristop z izkoriščanjem že obstoječih eno- in večjezičnih jezikovnih virov vključuje tudi visoko stopnjo avtomatizacije, kar raziskovalcem prihrani ogromno časa in finančnih sredstev. Zaradi modularne strukture *EuroWordNet* omogoča razmeroma preprosto vključevanje novih jezikov, kar so že izkoristili za številne jezike. Po zaključku projekta *EuroWordNet* sta nastali leksikalni zbirki za švedski in ruski jezik, rezultat projekta *BalkaNet* (Tufis in dr. 2004) pa so besedne mreže za bolgarski, grški, romunski, srbski in turški jezik.

Poleg prednosti prinaša razširitveni pristop tudi številne negativne posledice, med katerimi je nedvomno najpomembnejša prevelika odvisnost od leksikalne in konceptualne strukture izvornega jezika (največkrat angleškega). Če nismo dovolj pozorni, je lahko za določen jezik izdelana leksikalna zbirka arbitrarna in z dejansko organizacijo in leksikalizacijo konceptov v tem jeziku nima veliko skupnega (glej Vider 2004 in Wong 2004).

Pri tem pristopu naletimo na posebnosti, ki jih lahko razdelimo v dve skupini: **leksikalne vrzeli** (pojem, ki je v nekem jeziku izražen z leksikalno enoto, je v drugem mogoče izraziti samo s prosto kombinacijo besed) in **denotacijske razlike** (v ciljnem jeziku obstaja prevodna ustreznica pojma izvornega jezika, vendar je nekoliko splošnejša ali nekoliko bolj specifična). V obeh primerih leksikalni pojem v izhodiščnem jeziku nima sopomenske ustreznice v drugem jeziku (Bentivogli in dr. 2004).



Slika 1: *Klasifikacija prevodnih ustreznic (prirejeno po Bentivogli in dr. 2004).*

3.1 Izdelava slovenske besedne mreže z razširitvenim pristopom

Strog razširitveni pristop za gradnjo slovenske mreže s splošnim besediščem smo preizkusili na pomenskem polju |sorodstvo|. Iz *WordNeta* 2.0 smo prevzeli hierarhično drevo in ga z uporabo splošnih in strokovnih dvo- in enojezičnih slovarjev prevedli v slovenščino (glej **sliko 2**). V nadaljevanju pojasnjujemo rezultate in izpostavljamo največje probleme tovrstnega pristopa za izdelavo leksikalnih zbirk.

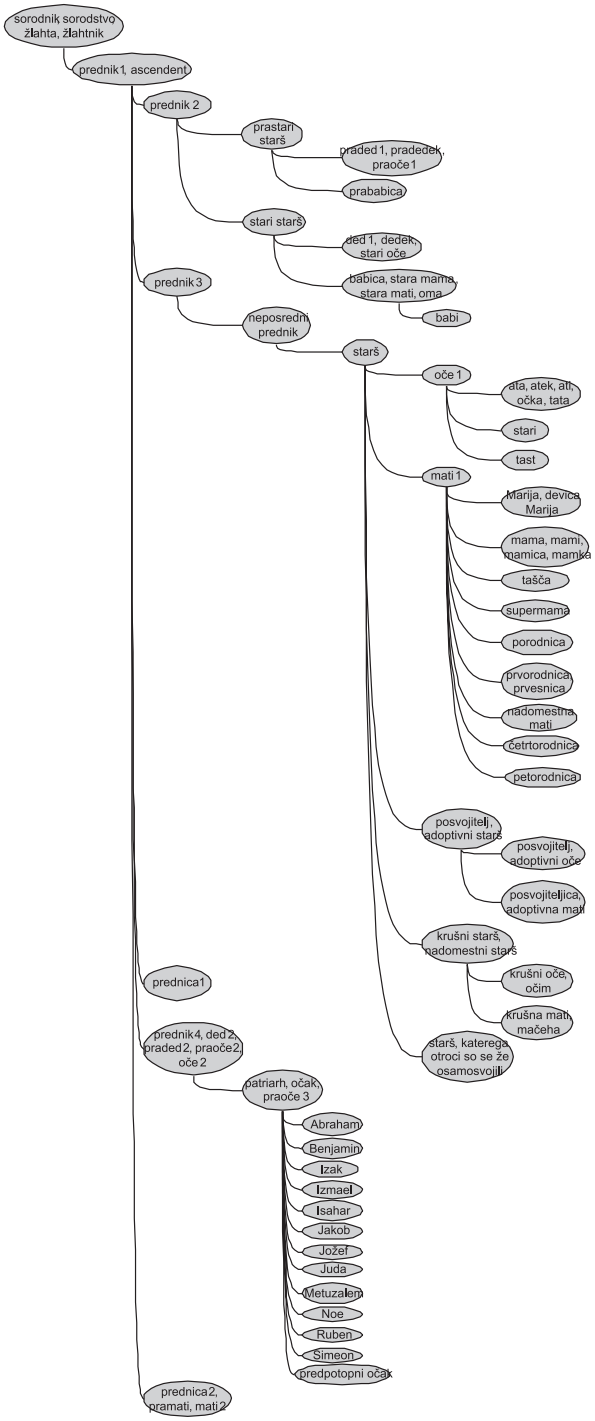
3.1.1 Kulturnospecifični pojmi

V prevzetem hierarhičnem drevesu so očitni sinseti z biblijsko vsebino: {Marija, devica Marija} (ang. {*Mary, Virgin Mary, The Virgin, Blessed Virgin, Madonna*}), {Abraham} (ang. {*Abraham, Ibrahim*}), {Benjamin} (ang. {*Benjamin*}), {Izak} (ang. {*Isaac*}), {Izmael} (ang. {*Ishmael*}), {Isahar} (ang. {*Isaachar*}), {Jakob} (ang. {*Jacob*}), {Jožef} (ang. {*Joseph*}), {Juda} (ang. {Judah}), {Metuzalem} (ang. {*Methuselah*}), {Noe} (ang. {*Noah*}), {Ruben} (ang. {*Rueben*}), {Simeon} (ang. {*Simeon*}) in {predpotopni očak} (ang. {*antediluvian, antediluvian patriarch*}).

Ti sinseti so problematični iz dveh razlogov: prvič, naš cilj je gradnja besedne mreže s splošnim besediščem, kamor našteti sinseti zagotovo ne sodijo in jih pri gradnji neodvisne leksikalne zbirke najverjetneje ne bi vključili. In drugič, vsebina teh sinsetov je versko motivirana in zato tudi kulturnospecifična. Če bi šlo za pomembne pojme iz drugih kultur, npr. hindujske, jih v slovensko leksikalno zbirko, ki jo želimo zgraditi, verjetno ne bi zajeli, tako kot druge kulture nimajo razlogov za vključevanje krščanskih vsebin. S tem se zastavlja pomembno vprašanje (ne)smiselnosti medkulturne in medjezikovne prenosljivosti pojmov.

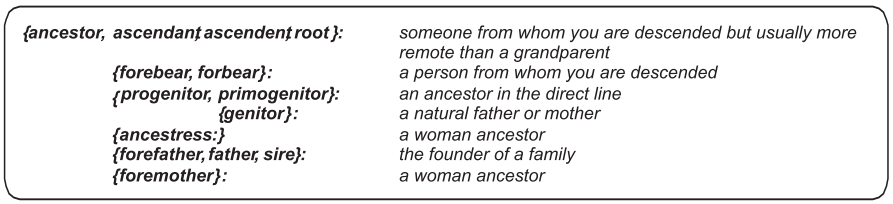
3.1.2 Denotacijske razlike

V primerjavi s slovenščino je v angleščini veliko več izrazov za pojem *prednik*, saj je 14 različnih izrazov zanj razdeljenih na sedem različnih sinsetov. S primerjavo vključenih definicij opazimo, da takšna drobitev pomenov ni nujno upravičena (glej **sliko 3**). Ne glede na to pri razširitvenem pristopu v slovenskem jeziku dobimo kar



Slika 2: *Rezultat razširitvenega pristopa.*

nekaj nesmiselnih hierarhičnih stopenj in sinsetov z enako vsebino. Do tega prihaja zaradi denotacijskih razlik med angleškimi in slovenskimi izrazi za pojem *prednik*, pri čemer je slovenska prevodna ustreznica največkrat splošnejša od angleškega izvirnika (npr. {prednik} za ang. {*forebear, forbear*} in {*progenitor, primogenitor*}). Tolikšna drobitev pomenov leksema *prednik* v slovenščini ni izražena, zato je pojavitev tega izraza na štirih različnih hierarhičnih ravneh v slovenščini povsem arbitrarna in zahteva nujen razmislek o alternativnih rešitvah.



Slika 3: *Hierarhična struktura angleških izrazov za pojem prednik, vključenih v* WordNet.

3.1.3 Leksikalne vrzeli

Leksikalne vrzeli se pojavijo, kadar je pojem v izvornem jeziku leksikaliziran, v ciljnem pa ga lahko izrazimo samo opisno, torej s prosto kombinacijo besed. V našem primeru se to pojavi pri sinsetih {predpotopni očak} (ang. {*antediluvian, antediluvian patriarch*}), {starši, katerih otroci so se že osamosvojili} (ang. {*empty nester*}), in {neposredni prednik} (ang. {*genitor*}). V povezavi z angleškim izrazom *empty nester* v slovenščini sicer poznamo besedno zvezo *sindrom praznega gnezda*, vendar za starše, ki se po osamosvojitvi otrok počutijo osamljeni in nesrečni zaradi izgube svoje družbene vloge, v korpusu nismo našli posebnega izraza. Podobno opazimo pri kolokaciji *neposredni prednik*, ki jo v korpusu *Fida* sicer najdemo, a med 28 pojavitvami nismo našli nobenega primera, ki bi izražal obravnavani pomen (ang. {*genitor*}). Tovrstni sinseti v leksikalni podatkovni zbirki nimajo praktične uporabne vrednosti, zato se zdi njihovo vključevanje v zbirko nesmiselno.

3.1.4 Semantične relacije

Kot smo že omenili, temelji razširitveni pristop na predpostavki, da relacija, ki povezuje sinseta v izvornem jeziku, povezuje tudi ekvivalentna sinseta v ciljnem jeziku. V nadaljevanju razdelka se podrobneje posvečamo posameznim semantičnim in leksikalnim relacijam in preverjamo smiselnost njihovega prenosa v slovenski jezik.

Sopomenskost

Leibniz besedi definira kot sopomenki, če zamenjava ene z drugo *nikoli* ne spremeni pomena stavka, v katerem je do zamenjave prišlo. Miller pa je definicijo razširil tako, da sta vsaj dva različna leksema sopomenki, če med njima velja besedilna zamenljivost, ki ne spremeni pomena stavka v *določenem* kontekstu (citirano po Fellbaum 1998: 24). Pri tem imamo v mislih ohranjanje denotativnega pomena, konotativnosti, pragmatičnega vidika (zvrstne, slogovne in siceršnje zaznamovanosti) pa ne upoštevamo. Čista pomenska prekrivnost je v jeziku zaradi načela ekonomičnosti zelo redka (Vidovič Muha 2000: 161). Sopomenskost oziroma sinonimija je horizontalni pojav, relacija pa je simetrična: če je *x* sopomenka besede *y*, je tudi *y* sopomenka besede *x*. Nadalje med sopomenkami velja načelo homogenosti: če sta besedi sopomenki, ju ne more povezovati nobena druga semantična relacija (Fellbaum 1998: 24–44).

Stavek	Primer	Odgovor
Če je *x*, potem je tudi *y*.	Če je avto, potem je tudi avtomobil.	da
Če je *y*, potem je tudi *x*.	Če je avtomobil, potem je tudi avto.	da

Tabela 1: *Hallidayevi identifikacijski stavki za določanje sopomenk.*

Nadpomenskost in podpomenskost

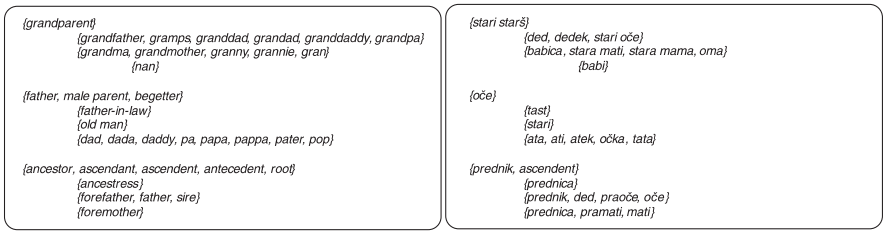
Nadpomenskost in podpomenskost (hiper- in hiponimija) sta temeljni urejevalni načeli leksikalnih zbirk in našega načina mišljenja. Nadpomenskost in podpomenskost sta inverzni relaciji: če je *y* neke vrste *x*, potem je *x* nadpomenka besede *y*, *y* pa podpomenka besede *x*. Uvrščevalne pomenske sestavine nastopajo v nadpomenski vlogi, razločevalne pa v podpomenski, zato sta obe relaciji asimetrični in tranzitivni (Muha 2000: 174–177). V referenčnem kontekstu podpomenko lahko zamenjamo z nadpomenko (specifično besedo zamenjamo s splošnejšo), nadpomenke s podpomenko pa ne (če jo že zamenjamo, gre za metonimijo). Za podpomenke je značilno, da podedujejo vse lastnosti svojih nadpomenk, relacija pa je resnično podpomenska šele takrat, kadar ima nadpomenka več kot eno podpomenko (Fellbaum 1998: 24–44). Med podpomenkami ne velja razmerje privativnosti (prisotnost oz. odsotnost določene lastnosti), temveč ekvipolentnosti (kohiponimija oz. sorednost) (Vidovič Muha 2000: 177).

Stavek	Primer	Odgovor
Če je *y*, potem je tudi *x*.	Če je nemški ovčar, potem je tudi pes.	da
Če je *x*, potem je tudi *y*.	Če je pes, potem je tudi nemški ovčar.	ne
Y je neke vrste *x*.	Nemški ovčar je neke vrste pes.	da
X je neke vrste *y*.	Pes je neke vrste nemški ovčar.	ne

Tabela 2: *Hallidayevi identifikacijski stavki za določanje nad- in podpomenk.*

V angleškem *WordNetu* gre za na nekonsistentno obravnavanje nad- in podpomenskosti. Primer za to so razmerja med sinseti v sliki 5. Medtem ko sinset za pojem *stari oče* kot sinonime obravnava tako nevtralne kot zaznamovane lekseme, so le-ti pri pojmu *oče* ločeni. Podobno opazimo pri pojmu *mati*. Prav tako neutemeljeno

pojem *stara mama* izraz *nan* uvršča kot podpomenko. Ker razširitveni model pri prevajanju besedne mreže v slovenščino vse relacije ohrani, se ohranijo tudi omenjene nekonsistentnosti.



Slika 4: *Nad- in podpomenske relacije v angleškem* WordNetu *s prevodnimi ustreznicami.*

Primer, pri katerem se zastavi vprašanje, ali se relacije med sinseti zares ohranijo tudi v ciljnem jeziku, je angleški sinset {*father-in-law*}, ki je podpomenka pojma *oče*. Vprašanje je, ali v slovenščini pojma *tast* ne bi namesto v pojem *oče* uvrstili med *svaštvene sorodnike*, saj tast ni krvni sorodnik. Enako velja za pojem *tašča* (ang. {*mother-in-law*}).

Naslednji velik problem nad- in podpomenskosti je nesistematično obravnavanje ženskih oblik samostalnikov, saj sta na primer *forefather* in *foremother* obravnavana kot kohiponima, *ancestress* pa kot podpomenka pojma *ancestor*.

Protipomenskost

Tudi protipomenskost oziroma antonimija je temeljno medleksemsko razmerje, ki ga izražajo leksikalnima nasprotja, kot so npr. *dobro – slabo*, *pravica – krivica* in *gor – dol*. Jasno je, da je protipomenskost simetrična relacija, natančno definicijo pa je zelo težko postaviti, saj pokriva številne pojave nasprotij (Fellbaum 1998: 24–44). Z logičnega vidika ločujemo med medleksemsko nasprotnostjo (npr. *dan – noč*) in dopolnjevalno protislovnostjo (npr. *moški – ženska*). Protipomenskost se za razliko od sopomenskosti, ki je po številu sopomenk teoretično neomejena, lahko pojavlja le v parih. Isti večpomenski leksem se lahko pojavlja v različnih protipomenskih parih. Logičnost nasprotja znotraj leksemsko pomenskih parov temelji na pripadnosti denotatov skupnemu pomenskemu polju, zato se protipomenki med seboj razlikujeta v eni ali več lastnostih in imata skupno nadpomenko (Vidovič Muha 2000: 169).

Stavek	Primer	Odgovor
X in *y* sta oba neke vrste *z*, vendar je *x* nasprotje *y*.	Moški in ženska sta oba neke v rste človek, vendar je moški nasprotje ženske.	da
Y in *x* sta oba neke vrste *z*, vendar je *y* nasprotje *x*.	Ženska in moški sta oba neke vrste človek, vendar je ženska nasprotje moškega.	da

Tabela 3: *Hallidayevi identifikacijski stavki za določanje protipomenk.*

V obravnavanem pomenskem polju je kodificirana dopolnjevalna protislovnost (*mati – oče*) in ni problematična.

3.2 Poskus izdelave jezikovnomotivirane slovenske besedne mreže

Iz zgornje analize besedne mreže, ki smo jo dobili s pomočjo razširitvenega pristopa, lahko ugotovimo, da že angleški *WordNet*, iz katerega smo izhajali, vsebuje številne nekonsistentne in vprašljive rešitve, ki jih je slovenska mreža zaradi metodologije izdelave podedovala. Poleg tega so se pojavili tudi številni problemi zaradi sociološko-kulturoloških, konceptualnih in leksikalnih razlik med jezikoma. Zato smo poskusili izdelati besedno mrežo v slovenščini, ki ne bi bila odvisna od drugih kultur in jezikov, temveč bi temeljila na dejanskih lastnostih slovenskega jezika.

Kandidate zanjo smo pridobili s pomočjo angleškega *WordNeta*, enojezičnega referenčnega korpusa (*Fida*),[8] dvojezičnih in enojezičnih slovarjev (*Veliki angleško-slovenski slovar*, *Angleško-slovenski* in *Slovensko angleški pravni slovar*, *Slovar slovenskega knjižnega jezika* in *Slovenski pravopis*) in strokovnih priročnikov (*Eurovoc*, *Družinsko pravo*). Zaradi obsežnosti in razvejanosti pomenskega polja smo se omejili samo na sorodnike v ravni črti nazaj (predniki), izpustili pa vse sorodnike v ravni črti naprej (potomci) in sorodnike v stranski črti (bratje, sestre, strici, tete, bratranci, sestrične). Skupno število vseh kandidatov je bilo 140. Sledilo je izločanje kandidatov, ki smo jih našli samo v enem slovarskem viru, npr. samo v *SSKJ*, v korpusu *Fida* pa ne (*starodavnik, zarodnik*). Ker smo želeli izdelati besedno mrežo splošnega besedišča, smo izločili tudi kandidate, ki so v *SSKJ* označeni s kvalifikatorji *star., zastar.* in *nar.* in se v *Fidi* v tem pomenu pojavijo manj kot petkrat (*čača, dedec, dedej, otec, papači, majka, maman, mamika, nona, sprednik, sorodovinec, žlahtovec*), in kandidate, ki pravzaprav ne sodijo v to pomensko polje (*detomorilka, družinski poglavar, varuh, varuhinja*). Prav tako smo izpustili kolokacije, ki so sicer imele precej zadetkov v korpusu, vendar je njihov pomen bolj opisne kot kvalifikatorske narave in ga nismo mogli natančno določiti niti s pomočjo konkordanc niti z uporabo drugih virov (*bližnja sorodnica, bližnji sorodnik, bližnje sorodstvo, daljna sorodnica, daljni sorodnik, daljno sorodstvo, davni sorodnik, najožja sorodnica, najožji sorodnik, najožje sorodstvo, ožja sorodnica, ožji sorodnik, ožje sorodstvo, širše sorodstvo*).

[8] <http://www.fida.net> [15. 6. 2005].

adoptivna mati	krvna sorodnica	očim	samohranilka
adoptivni oče	krvni sorodnik	očka	skrbnica
adoptivni starši	krvno sorodstvo	oma	skrbnik
ascendent	lateralni sorodnik	otec	sorodnica
ascendentka	mačeha	ožja sorodnica	sorodnik
ata	majka	ožje sorodstvo	sorodnik po svaštvu
ate	mama	ožji sorodnik	sorodnik v svaštvu
atej	maman	papa	sorodovinec
atek	mami	papaček	sorodstvo
ati	mamica	papači	sprednik
babi	mamika	porodnica	stara mama
babica	mamka	posvojitelj	stara mati
biološka mati	mat	posvojiteljica	stari ata
biološki oče	mati	prababica	stari ate
biološki starši	mati samohranilka	praded	stari čača
bližnja sorodnica	matka	pradedek	stari foter
bližnji sorodnik	nadomestna mati	praoče	stari fotr
bližnje sorodstvo	nadomestna roditeljica	prastari starši	stari oče
čača	nadomestni oče	prastarši	stari starši
daljna sorodnica	nadomestni roditelj	prava mati	starodavnik
daljni sorodnik	nadomestni starši	pravi oče	starši
daljno sorodstvo	najožja sorodnica	pravi starši	svaštveni sorodnik
ded	najožji sorodnik	prednica	svaštvo
dedec	najožje sorodstvo	prednik	širše sorodstvo
dedej	naravna mati	prvesnica	tast
dedek	naravni oče	prvorodnica	tašča
dedi	naravni starši	rejnica	tata
detomorilka	nezakonska mati	rejnik	varuh
družinski poglavar	nezakonski oče	roditlej	varuhinja
foter	nona	roditeljica	zarodnik
fotr	nono	rodna mati	žlahta
krušna mati	oča	rodni oče	žlahtnica
krušni oče	oče	rodnica	žlahtnik
krušni starši	oči	rodnik	žlahtovec

Slika 5: *Seznam kandidatov za slovensko semantično mrežo.*

3.2.1 Uporaba korpusa pri gradnji semantičnih leksikonov

Ostalo nam je nekaj manj kot sto kandidatov, med katerimi smo s pomočjo konkordanc ali razširjenega sobesedila v korpusu *Fida* skušali najti semantične in leksikalne relacije. V nadaljevanju predstavljamo nekaj primerov, na podlagi katerih smo lahko sklepali na relacije med posameznimi pojmi.

Primeri na naslednji strani **(tabela 4)** upravičujejo vključevanje leksemov *ata*, *atek*, *ati*, *oče*, *očka* in *tata* v isti sinset. Iz najdenih primerov je prav tako mogoče ugotoviti protipomenski odnos med leksemoma *mama – ata* ter dokaze, da ima lahko *ata* dva različna pomena. Vidimo lahko, da je enkrat vzpostavljena razlika med *ata* in *stari ata* ter *dedek*, drugič pa med istimi besedami opazimo sopomenskost. Zato smo v slovensko besedno mrežo leksem *ata* vključili dvakrat, enkrat v sinset {ata1, atek, ati, oče, očka, tata}, drugič pa v sinset {ata2, ded, dedek, stari ata, stari oče}. Posamezna pomena smo med seboj ločili s številkama; številko 1 ima pomen, ki se v korpusu pojavi večkrat.

Sobesedilo	Relacija
ata v pomenu oče	
Jezus se je na svojega Očeta obračal z besedo »Abba«, kar bi lahko prevedli z »**očka**, **ata**, **tata**, **papa**«.	sopomenskost: očka – ata – tata – papa – oče
ljubkovalni naziv za **očeta**, **atek**	sopomenskost: očka – atek
Kaj delaš? Si še vedno v Osijeku? Pa tvoja **mama** in **ata**?	protipomenskost: ata – mama
Ob zadnjem slovesu od dragega moža, **ata** in **starega ata**	razlikovanje med ata in stari ata
Našega ljubega moža, **ata** in **dedka** bomo na zadnjo pot pospremili /.../	razlikovanje med ata in dedek
ata v pomenu stari oče, dedek	
Ata Janez (**dedek**) ga je kar dobro naučil.	sopomenskost: ata – dedek
16. februarja bosta minili dve leti žalosti, kar nas je za vedno zapustil dragi mož, **ati** in **ata.**	razlikovanje med ati in ata
Mama Ančka, kakor smo jo klicali vnuki je bila leto mlajša od **ata** Herberta. S Herbertom sta jo popravila in preuredila. Prav dosti o mami ne vem /.../ **Oče** je pripovedoval, da je bila velikokrat nergava	razlikovanje med ata in oče
Ob nenadni, boleči izgubi našega ljubega moža, **očeta**, **ata**, sina in brata	razlikovanje med ata in oče

Tabela 4: *Primeri relacij, najdenih v korpusu (vir:* Fida*).*

Na enak način dobimo rezultate za večino ostalih pojmov. Kot problematična pa se je pokazala kolokacija *nadomestna mati*, za katero ni povsem jasno, ali gre za žensko, ki namesto biološke matere prevzame skrb za otroka (ang. *foster mother*), ali za biološko mati, ki rodi otroka za neplodni par (ang. *surrogate mother*). S poizvedbo v korpusu dobimo naslednje rezultate:

pomen	št. poj.	sobesedilo
nadomestna mati = krušna mati	10	Pod določenimi pogoji se **rejništvo nadomestnim materam** lahko prizna tudi kot poklic /…/.
nadomestna mati = biološka mati	5	**Zarodke vsadijo v ma ternice** drugih, **nadomestnih mater**, ki prejemajo ustrezne hormone za vzdrževanje nosečnosti.
nadomestni oče = krušni oče	4	**Očeta** v času vojne praktično ne videva, zato **nadomestnega očeta** najde v dedku Gearu /…/.
nadomestni starši = krušni starši	5	**Skupina rejnic in rejnikov** – **nadomestnih staršev** torej, se srečuje v Piranu vsako prvo sredo v mesecu

Tabela 5: *Rezultati poizvedbe po kolokaciji »nadomestna mati« (vir:* Fida*).*

Glede na korpus sta v slovenščini prisotna oba pomena, čeprav je prvi pogostejši. Če iščemo še kolokaciji *nadomestni oče* in *nadomestni starši*, ugotovimo, da je v korpusu zastopan samo prvi pomen, zato smo v slovensko besedno mrežo vključili vse tri kolokacije glede na prvi pomen, drugi pomen pa smo upoštevali samo za kolokacijo *nadomestna mati*.

Največje odstopanje od mreže, dobljene z razširitvenim pristopom, pa je v odmiku od drobljenja pomenov pojma *prednik*. V tem pomenskem polju smo upoštevali samo pomen leksema *prednik*, ki označuje sorodnike v ravni črti nazaj, ne pa človekovih

oddaljenih prednikov (kot so Slovani, neandertalec ipd.). V isti sinset smo vključili še pravni izraz *ascendent* in ženski obliki obeh samostalnikov *prednica* in *ascendentka*.

Pojem *starš* smo razdelili na dva pomena, in sicer glede na to, ali gre za biološke ali za nadomestne starše. Če bi se držali strukture iz angleškega *WordNeta*, bi bila pojma *krušni starš* in *posvojitelj* podpomenki pojma *starš*. To pa ob upoštevanju dejstva, da podpomenka podeduje vse lastnosti svojih nadpomenk, ni mogoče, saj bi v tem primeru trdili, da so krušni starši in posvojitelji otrokovi predniki in njegovi krvni sorodniki.

Različne izraze za porodnice smo obravnavali kot podpomenke sinseta {porodnica}, pri čemer smo izpustili izraza *četrtorodnica* in *petorodnica*, ker menimo, da termina s tako ozkega področja ne sodita v besedno mrežo splošnega besedišča, ki smo jo začeli zgraditi za začetek. Vsekakor pa ju je po potrebi mogoče kadar koli dodati, in sicer kod podpomenki sinseta {mnogorodnica}.

3.2.2 Grafična predstavitev rezultatov

Rezultate, dobljene s pomočjo korpusa, smo grafično predstavili v obliki hierarhičnega drevesa. Črtkana polja označujejo veje pomenskega polja, ki jih v tem prispevku zaradi preobsežnosti problema nismo obravnavali.

Elipse ponazarjajo posamezne sinsete. Ženske oblike samostalnikov, ki smo jih razvrstili v isti sinset kot njihove moške ustreznice, bi v leksikalni zbirki opremili z oznako *ženska oblika*. Tudi besedam, ki pripadajo različnim registrom, a jih kljub temu obravnavamo kot sopomenke nevtralnim izrazom, bi bilo treba dodati oznako (npr. *pogovorno, ljubkovalno, knjižno, medicinsko, pravno*).

Pri grafični predstavitvi rezultatov naletimo na težave pri nad- in podpomenskem odnosu med sinsetoma {starš1} → {ata1, oče, ate, atek, atej, ati, oča, oče, očka, papa, tata} in sinsetoma {biološki roditelj, biološki starš, naravni roditelj, naravni starš pravi roditelj, pravi starš, rodni starš} → {biološki oče, naravni oče, pravi oče, roditelj, rodnik, rodni oče}, saj je poleg teh zaznati tudi povezavo med sinsetoma {ata1, oče, ate, atek, atej, ati, oča, oče, očka, papa, tata} → {biološki roditelj, biološki starš, naravni roditelj, naravni starš pravi roditelj, pravi starš, rodni starš}. V trenutku, ko ima isti sinset več kot eno nadpomenko, se drevesna struktura spremeni v mrežo, kar je problematično za računalniške aplikacije. Zato smo bili prisiljeni upoštevati samo prvi dve relaciji, tretje pa nismo eksplicitno izrazili. Enako rešitev smo uporabili za pojem *mati*.

Ta težava dokazuje, da je pri formalizaciji naravnega jezika nujno potrebna določena mera posploševanja in odmika od leksikona v kognitivno-nevrološkem smislu, zaradi česar v modele mentalnega leksikona ne moremo zajeti čisto vseh prvin naravnega jezika. Gradnja besednih mrež pri tem ni nobena izjema, česar se moramo pri delu z njimi vselej zavedati.

Nadpomenskost in podpomenskost med besedami ponazarjajo neprekinjene črte. Čeprav je iz konkordanc mogoče sklepati tudi o protipomenskih relacijah, jih zara-

di boljše preglednosti nismo vključili v drevesno strukturo, v leksikalni zbirki pa bi jih brez dvoma morali upoštevati. Identificirali smo tako dopolnjevalno protislovnost, npr. *dedek – babica*, *oče – mati* in medleksemsko nasprotnost, npr. *krvni sorodnik – nekrvni sorodnik*, *prvorodnica – mnogorodnica*.

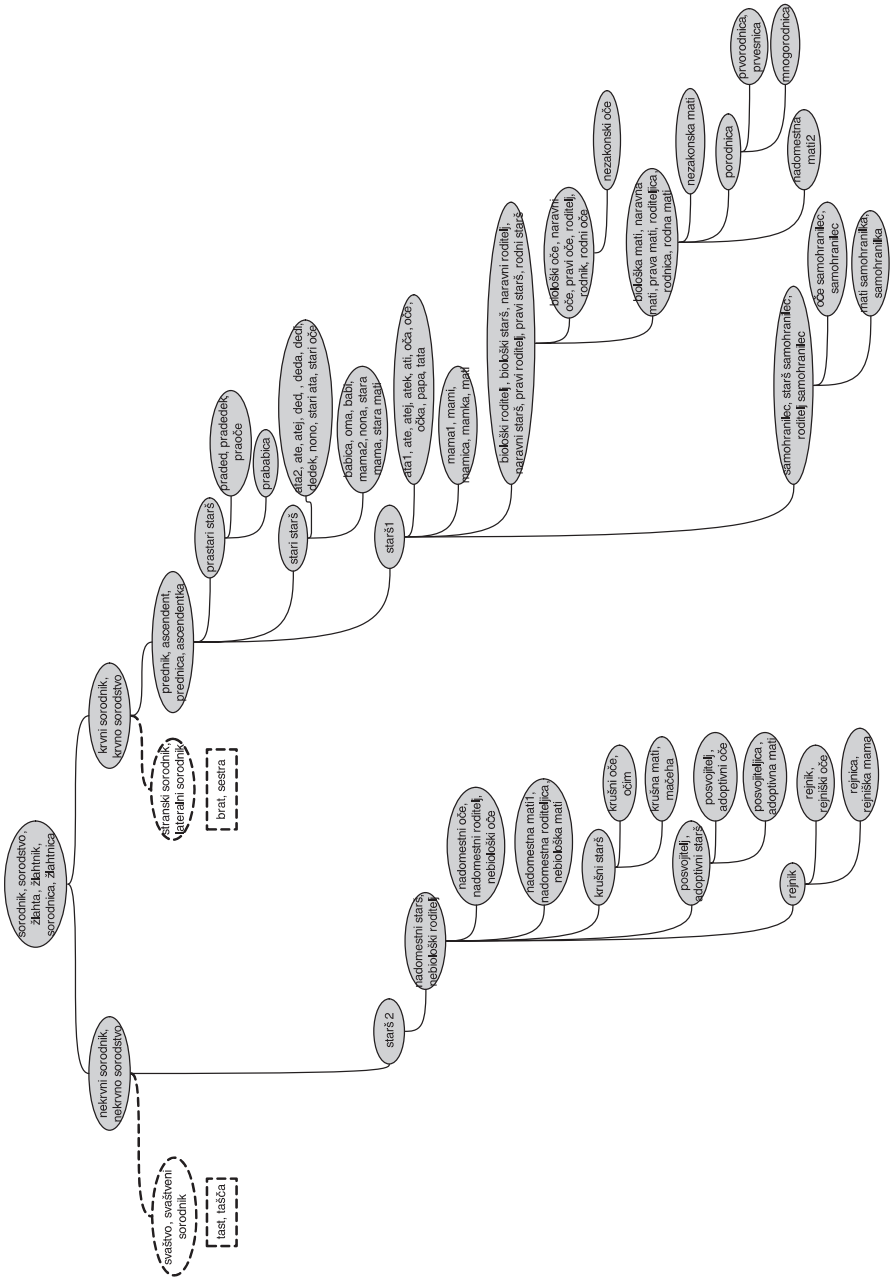
4 Zaključek

Besedni mreži, ki smo ju izdelali s pomočjo razširitvenega in jezikovnomotiviranega pristopa se precej razlikujeta, kar zastavlja pomembno vprašanje, kateri od njiju bolj verodostojno odraža dejanska slovenska pomenska razmerja med posameznimi izrazi, saj je zelo poemembno, da besedna mreža vsebuje pojme, ki so v slovenskem jeziku zares leksikalizirani, in povezave med njimi, ki niso zgolj arbitrarne.

Za natančnejšo presojo o tem bi bilo kvaliteto rezultatov potrebno še izmeriti z ustrezno metodo, na prvi pogled pa se zdi, da je ustreznejši jezikovnomotivirani pristop, pri katerem smo uporabili enojezični referenčni korpus *Fida* in ostale eno- in večjezične jezikovne vire. Ker smo pri tem pristopu bolj upoštevali dejansko rabo, se v besedni mreži ne pojavljajo neleksikalizirani pojmi, leksikalizirani pa so urejeni v skladu z njihovim dejanskim pojmovanjem v slovenskem jeziku in kulturi.

Postopek pridobivanja informacij za izdelavo besedne mreže bi bil veliko hitrejši in enostavnejši, če bi imeli na razpolago ustrezen vzporedni angleško-slovenski korpus, kjer bi lahko preverili prevodne ustreznice za posamezne izraze. Zelo koristno bi bilo tudi, če bi bil korpus *Fida* že razdvoumljen, kar je sicer eden od možnih področij uporabe izdelane besedne mreže za slovenski jezik, vendar je razdvoumljanje mogoče tudi z drugimi metodami. V tem primeru bi bilo iskanje želenih primerov veliko hitrejše, kar se pozna predvsem pri zelo pogostih večpomenskih izrazih, kot je na primer *babica*. Iskanje po lemi prikaže 2619 zadetkov, vendar niso vsi relevantni. Poleg pomena *stara mati* so namreč med rezultati še drugi: *stara ženska*, *pomočnica pri porodih*, *morska riba* in *priprava za zapenjanje*. Podatki o številu pojavitev so v tem primeru nerelevantni, saj ne vemo, kolikokrat je med njimi zastopan pomen *stara mati*, pregled vsakega zadetka posebej pa zelo zamuden in dopušča veliko možnosti za napake.

Tudi z alternativnim pristopom nam ni uspelo izdelati besedne mreže, ki bi popolnoma ustrezala dejanski jezikovni rabi, vendar smo z njim kljub temu dobili veliko boljše rezultate, kot jih je dal razširitveni pristop. Priznati pa je treba, da je tak način izdelave besednih mrež zelo dolgotrajen in drag postopek, kar je velika slabost v primerjavi z razširitvenim pristopom, ki omogoča visoko stopnjo avtomatizacije. Vendar je poleg cene pomemben dejavnik tudi uporabnost izdelanih aplikacij. Znano je, da računalniško jezikoslovje pogosto posega po delnih (in udobnih) rešitvah, s katerimi (tradicionalno) jezikoslovje nikoli ni povsem zadovoljno in ki se na prvi pogled zdijo veliko bolj uporabne, kot tudi dejansko so (Čermák 2002: 276). Zato za nastanek resnično uporabne leksikalne podatkovne zbirke za slovenski jezik nujno potrebujemo sodelovanje obeh disciplin, ki bo z vmesnim pristopom obrodilo cenovno še sprejemljivo, a vsestransko uporabno rešitev.



Slika 6: *Slovensko hierarhično drevo pomenskega polja* |sorodstvo|.

Literatura

Aitchison, Jean, 2004: *Words in the Mind. An Introduction to the Mental Lexicon*. Oxford: Blackwell Publishing (3rd edition).

Bentivogli, Luisa in Pianta, Emanuele, 2004: *Looking for Lexical Gaps*. V zborniku mednarodnega kongresa *Euralex-2000*. <http://multiwordnet.itc.it/paper/wordnet-euralex2000.pdf> [13. 6. 2005].

Čermák, František, 2002: Today's corpus linguistics. Some open questions. Teubert, Wolfgang (ur.): *International Journal of Corpus Linguistics* 7/2. 265–282.

Fellbaum, Christine (ur.), 1998: *WordNet. An Electronic Lexical Database.* Cambridge: MIT Press.

Noy, Natalia, 2003: Ontologies. Farghaly, Ali (ur.): *Handbook for Language Engineers*. Stanford: CSLI Publications.

Tufis, Dan idr., 2004: BalkaNet: Aims, Methods, Results and Perspectives. *A General Overview*. Dascalu, Dan (ur.): *Romanian Journal of Information Science and Technology* 7/1–2. 9–43.

Vider, Kadri, 2004: Concerning the Difference Between a Conception and its Application in the Case of the Estonian WordNet. Sojka, Petr idr. (ur.): *Proceedings of the Global WordNet Conference*. 285–290.

Vidovič Muha, Ada, 2000: *Slovensko leksikalno pomenoslovje: govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Vossen, Piek (ur.), 1998: *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht: Kluwer Academic Press.

Vossen, Piek (ed.), 2005: *EuroWordNet. General Document (final version)*. <http://www.hum.uva.nl/~ewn> [13. 6. 2005].

Zupančič, Karel, 1999: *Družinsko pravo*. Ljubljana: Uradni list Republike Slovenije.

Wong, Shun Ha Sylvia, 2004: Fighting arbitrariness in *WordNet*-like lexical databases. A natural language motivated remedy. *Proceedings of the Second Global WordNet Conference 2004*. Brno: Masaryk University. 234–241.